Name:_____ Date: _____

A2CC: Sample Proportions

Are cell phones a distraction?

A study released by the Centre for Economic Performance at the London School of Economics looked at 91 schools in four cities in England, where 90 percent of teenagers own a mobile phone. The study found that test scores were 6.41 percent higher where cellphone use is prohibited. In another study, more than half of students said they were distracted when other students used their devices.

Suppose we wanted to find the proportion of Roslyn high school students who believe that cell phones, Ipads, and other digital devices are a distraction in the classroom. Do I need to ask EVERY student to get an accurate measure? What if the population of the school was so large that it would be nearly impossible to ask each and every student?

Now, suppose that the true proportion of Roslyn high school students who believe that digital technology is a distraction in the classroom is 70%. That is, $p = .7$. Let's also assume that we don't know this value but would like to estimate it. When we take a sample from the population and find the proportion of the sample, $\hat{p}$, that believe technology is a distraction, is this a reasonable estimate for the true value of p?

Statistics can answer this question for us. But, first, we need to understand how statistics behave.

The statistic, $\hat{p}$. (read as "p-hat")

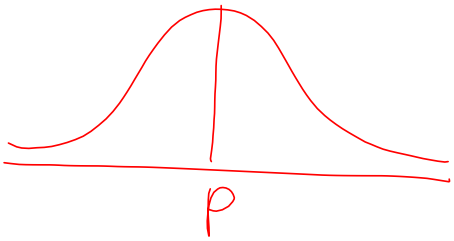
Suppose Mary takes a random sample of 50 Roslyn students and calculates $\hat{p}$, the proportion of HER sample that finds technology a distraction. Is it likely that her value is $\hat{p} = .7$? Not really

If Bill completes his own survey of 50 students, should we expect his value of $\hat{p}$ to be the same as Mary's?

No because there is sample variability

Suppose it were possible to take EVERY possible sample of size 50 and compute $\hat{p}$ for each one. All of these $\hat{p}$'s would create their own statistical distribution. It has been proven in advanced statistics that the distribution of $\hat{p}$'s is **NORMAL** with a mean, $\mu_{\hat{p}} = p$, and a standard deviation, $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$. standard deviation

(The discussion is not quite that simple and there are some other conditions that we have to verify, but we will leave that discussion for AP Statistics.)



n

Exercise # 1: Mary takes a random sample of 50 Roslyn high school students. Suppose the true value of $p = .7$ (but remains unknown). What is the probability that Mary finds that at least 76% of her sample think technology is a distraction?

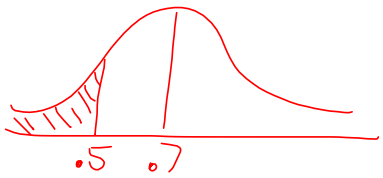
lower: .76
upper: 10000
μ: .7
σ =

$\sigma = \sqrt{\frac{.7(1-.7)}{50}}$

$P(\hat{p} \geq .76) = .177$ or 17.7%

Exercise # 2: Bill takes a random sample of 50 Roslyn high school students. Suppose the true value of $p = .7$ (but remains unknown). What is the probability that Bill finds that less than half of his sample think technology is a distraction?

.5



lower: -10000
upper: .5
μ: .7

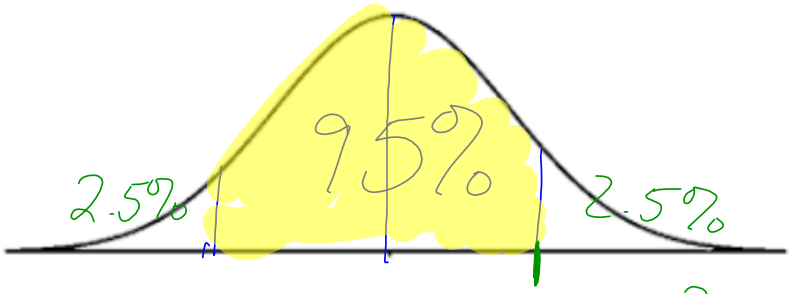
$\sigma = \sqrt{\frac{.7(.3)}{50}}$

$P(\hat{p} < .5) = .001$
.1%

In both of the previous examples, it was assumed that the true value of $p = .7$. However, if we KNEW that $p = .7$, we wouldn't need to gather data in the first place. This is where statistics come in handy. Since we know the characteristics of the distribution of $\hat{p}$, we can use these characteristics to make inferences about the true UNKNOWN value of p.

Before we can do this, we need to consider one more idea.

Suppose we want to "target" the middle 95% of a normal distribution, how many standard deviations below and above the mean must we go?



z-score?

$z = 1.96$

Accept ~~R~~ $p \pm 2\sqrt{\frac{p(1-p)}{n}}$

Confidence Intervals

So, the middle 95% of the distribution of $\hat{p}'s$ would be represented by $p \pm 1.96\sqrt{\frac{p(1-p)}{n}}$. However, in practice, we don't know p. Since $\hat{p}$ will be close to p in large samples, we replace p in the formula with $\hat{p}$ to create $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. What we've created is a 95% confidence interval for p. We use the term "95% confident" because if we were to construct a confidence interval for each $\hat{p}$ from EVERY possible sample, 95% of the confidence intervals created would contain the true value of p. What does the term 95% confident NOT mean? Most novices think that there is a 95% CHANCE that p is in the confidence interval. This is an INCORRECT interpretation of 95% confidence. The value of p is either in the interval or not. The probability that p is in the interval is either 1 or 0 not .95. We use the term 95% confident to describe the PROCEDURE that was used to create the interval.

Exercise #3: Revisiting Mary's sample from exercise #1. Using her sample, Mary calculated $\hat{p}=.76$. Create a 95% confidence interval for p.

$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$.76 \pm 1.96\sqrt{\frac{.76(.24)}{50}}$

So, we say …

95% confident the true value is between .642 and .878

$(.642, .878)$

Every confidence interval is of the form: estimate $\pm$ margin of error.

Other common confidence intervals…

A 90% confidence interval for p is $\hat{p} \pm 1.645\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

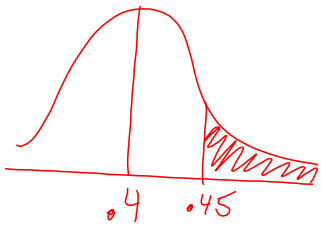
A 99% confidence interval for p is $\hat{p} \pm 2.576\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

The most commonly used is the 95% confidence interval.

$p = .4$

Exercise # 4: The Census Bureau reports that 40% of the 50,000 families in a particular region have more than one smart television in their household. What is the probability that a random sample of size 100 will have at least 45 % of the households that contain a smart television?

n



lower: .45
upper: 10000
$\mu = .4$
$\sigma = \sqrt{\frac{.4(.6)}{100}}$

$P(\hat{p} \geq .45) = .154$

15.4%

Exercise # 5: A large high school has approximately 1200 seniors. The school administration claims that 87% of its graduates are accepted into colleges. If a random sample of 80 graduates is taken, what is the probability that at most 64 of them are accepted into colleges?

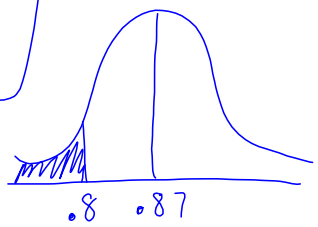
n

$p = .87$

$P(\hat{p} \leq \frac{64}{80}) = .031$

or 3.1%

$\frac{64}{80} = .8$



lower: -10000
upper: .8
$\mu = .87$
$\sigma = \sqrt{\frac{.87(.13)}{80}}$

Exercise # 6: Suppose that a random sample of 100 high school seniors in a particular city is taken, and it is found that 15% of the students favor the ban on prayer in public schools. Someone argues that the true proportion of seniors that favor the ban is 25%. What do your findings say about this?

$p = .15$

Find a confidence interval

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$.15 \pm 1.96\sqrt{\frac{.15(.85)}{100}}$$

$(.08, .22)$

We are 95% confident that the true proportion is between 8% and 22%.

Therefore we reject the idea that the true value is 25%.

A2CC: Sample Means

Same idea as sample proportions but with a different ⊙ equation.

Now, we are interested in estimating μ, the true population mean for a population of quantitative data with a known standard deviation, σ.

In order to estimate μ, we obviously take a random sample and calculate the mean of our sample, $\bar{x}$. But, once again, we are faced with some questions. Is it likely that my $\bar{x}$ is equal to the value of μ? Will my friend, who conducts his own sample, get the same value for his $\bar{x}$?

The statistic, $\bar{x}$. (read as "x-bar")

Suppose it were possible to take EVERY possible sample of size n and compute $\bar{x}$ for each sample. All of these $\bar{x}'s$ would create their own statistical distribution. It has been proven in advanced statistics that if the original distribution of values is normal, than the distribution of $\bar{x}'s$ is **NORMAL** with a mean, $\mu_{\bar{x}} = \mu$, and a standard deviation, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, where n is the size of the sample.

More impressively, a theorem in statistics, known as the Central Limit Theorem, states that even if the original distribution of values is non-normal, like salaries, that the distribution of $\bar{x}'s$ is close to a **NORMAL** distribution with a mean, $\mu_{\bar{x}} = \mu$, and a standard deviation, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, when the size of the sample is large.

(The discussion is not quite that simple and there are some other conditions that we have to verify, but we will leave that discussion for AP Statistics.)

Exercise # 1: The mean height of adult American males is 177 cm with a standard deviation of 7.3 cm. What is the standard deviation of the distribution of samples means from this population with a sample size of 50?

(1) 0.15 (2) 1.03 (3) 3.54 (4) 4.72

Confidence Intervals

Just as we did with sample proportions, we can create confidence intervals for a population mean.

The middle 95% of the distribution of $\bar{x}'s$ would be represented by $\mu \pm 1.96\frac{\sigma}{\sqrt{n}}$, where n is the size of the sample. However, in practice, we don't know μ. Since $\bar{x}$ will be close to μ in large samples, we replace μ in the formula with $\bar{x}$ to create $\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$. What we've created is a 95% confidence interval for μ. Just as before, we use the term 95% confident to describe the PROCEDURE that was used to create the interval.

Exercise # 5: A sample of 50 ripe oranges was taken from a large orchard in order to estimate the mean weight of a ripe orange. The sample mean was 212 grams and the sample standard deviation was 34 grams. Find a 95% confidence interval for the mean weight of a ripe orange in the orchard.

So, we say …

Every confidence interval is of the form: estimate $\pm$ margin of error.

Other common confidence intervals…

A 90% confidence interval for μ is $\bar{x} \pm 1.645\frac{\sigma}{\sqrt{n}}$.

A 99% confidence interval for μ is $\bar{x} \pm 2.576\frac{\sigma}{\sqrt{n}}$.

The most commonly used is the 95% confidence interval.

Exercise # 6: A random sample of 40 packages of light bulbs indicated that the mean number of defective bulbs in each package was .79 with a standard deviation of .2. Each package contains 4 bulbs.

a) Define the parameter of interest

b) Construct a 90% confidence interval based on these data. Include a statement.

Name: Key Date:

'2CCH: Regression homework

1. Which of the following linear equations would best fit the data set shown below?

(1) $y = 2.4x + 18.7$ (3) $y = -1.6x + 27.2$

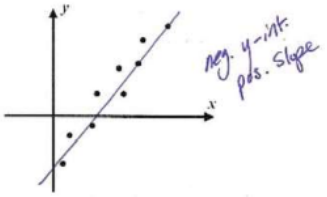
(2) $y = 0.8x + 18.1$ (4) $y = -1.9x - 15.6$

x	2	5	9	15
y	26	17	12	4

2. A scatter plot is shown below. Which of the following *could* be the equation of the best fit line for the data set?

(1) $y = 1.8x - 3.2$ (3) $y = -2.9x + 8.3$

(2) $y = -3.5x - 12.4$ (4) $y = 6.5x + 3.9$



3. A line of best fit was created for a data set that only included values of x on the interval $12 \le x \le 52$. For which of the following values of x would using this model represent extrapolation?

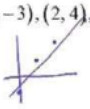
(1) $x = 26$ (3) $x = 14$

(2) $x = 50$ (4) $x = 6$

4. Which of the following is true about the line of best fit for the data set given in roster form below?

(1) It has a positive slope and negative y-intercept.
(2) It has both a positive slope and y-intercept.
(3) It has both a negative slope and y-intercept.
(4) It has a negative slope and positive y-intercept.

$\{(0,-3),(2,4),(6,10),(15,12)\}$



APPLICATIONS

5. An agronomist is studying the height of a corn plants as a function of the number of days since the corn germinated (appeared above the ground). Based on the following data, use your calculator to determine the best fit line in $y = ax + b$ form. Round all coefficients to the nearest *tenth*.

Time, x (days)	3	8	12	20	28	32	40
Height, y (inches)	2.5	4.5	6.2	9.3	12.9	14.4	16.8

$y = .4x + 1.4$

6. Heavier cars typically get worse gas mileage (their miles per gallon) than lighter cars. The table below gives the weight versus the highway gas mileage for seven vehicles.

Vehicle Weight (thousands of pounds)	2.5	2.9	3.1	3.0	4.2	6.6	3.4
Gas Mileage (miles per gallon)	34	36	31	29	23	12	26

(a) Determine the best fit linear equation, in $y = ax + b$ form, for this data set. Round all coefficients to the nearest tenth.

$y = -5.5x + 47.6$

(b) Using your model from part (a), determine the gas mileage, to the nearest mile per gallon, for a vehicle that weighs 3500 pounds.

$y(3.5) = 28$

(c) Is the prediction you made in (b) and example of interpolation or extrapolation? Explain.

domain: $2.5 \le x \le 6.6$

3.5 is inside domain

(d) What is the value of the correlation coefficient to the nearest *hundredth*? Why is it negative?

$r = -.95$

negative assoc. between weight + mileage

7. The superintendent of the Clarksville Central School District is attempting to predict the growth in student population in the coming years. The table below gives the population for her district for selected years.

Year	1990	1992	1995	1997	2002	2005
District Population	3520	3605	3771	3860	4135	4285

(a) Find the equation for the line of best fit, in $y = ax + b$ form, where x represents the years *since* 1990 and y represents the district's population. Round all coefficients to the nearest *hundredth*.

$y = 51.61x + 3509.98$

(b) Use your model from part (a) to predict the district's population in the year 2020. Round your answer to the nearest whole number.

$y(30) = 5058$

(c) What are the units of the slope of this linear model?

students per year

(d) What does the slope of this model represent? Think about your answer to part (c).

Each year, the student population is increasing by 51.61 students

8. Rabbits were accidently introduced to an island where their population is growing rapidly. Biologists studying the rabbits have periodically recorded their population since they were introduced to the island. The data they took is shown below.

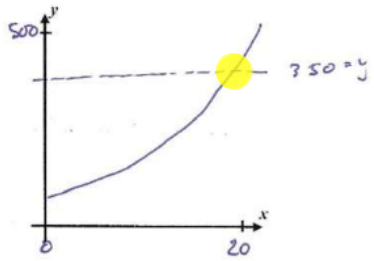
Years Since Introduction, x	2	5	7	11	15
Population of Rabbits, y	75	100	112	205	290

expreg!

(a) Determine an exponential regression equation, in the form $y = a \cdot b^x$, that models this data. Round a to the *tenth* and b to the *hundredth*.

$y = 58.2(1.11)^x$

(b) Sketch a graph of the rabbit population below on the axes provided for $0 \le x \le 20$. Label your graphing window and your y-intercept.



(c) Based on your model in part (a), by what percent is the rabbit population growing each year?

11% per year

(d) Graphically determine, to the nearest *tenth* of a year, when the rabbit population will reach 350.

16.6 years

9. The infiltration rate of a soil is the number of inches or water per hour it can absorb. Hydrologists studied one particular soil and found its infiltration rate decreases exponentially as a rainfall continues.

Time, t (hours)	0	1.5	3.0	4.5	6.0
Infiltration Rate, I (inches per hour)	5.3	3.1	2.4	1.6	0.7

Create an exponential model that best fits this data set. Round parameters to the nearest *hundredth*. Use your model to algebraically determine the time until the rate reaches 0.25 inches per hour. Round your answer to the nearest *tenth* of an hour. Use a logarithm in the process of your algebraic solution.

$y = 5.47(.73)^x$

$.25 = 5.47(.73)^x$

$.0457038391 = .73^x$

$\log_{.73} .0457038391 = x$

$9.8 = x$

10. The soil's temperature beneath the ground varies in a periodic manner. A temperature probe was left 3 feet underground and recorded the temperature as a function of the number of days since January 1st ($x = 0$ The temperatures for 14 days throughout the year are shown below.

Day	5	36	57	94	127	153	192
Temp (°F)	41	37	36	40	48	64	68
Day	226	241	262	289	305	337	356
Temp (°F)	66	61	58	49	44	42	40

(a) Find a best fit sinusoidal function for this data set in the form $y = a\sin(bx+c)+d$. Round all parameters to the nearest *hundredth*. Recall that some calculators require that you input the period on this correlation (365 days).

$y = 15.21\sin(.02x - 2.59) + 51.99$

(b) Based on your model from (a) what are the highest and lowest temperature reached in the soil?

$36.78^{\circ}F \le T \le 67.2^{\circ}F$

(c) What is the average soil temperature?

49.973 using calculus

poor question

51.99 ← midline

(d) If the root of a particular plant species will only thrive when the soil temperature is above 50 °F, graphically determine the interval of days over which the plant will thrive.

$120.5 < x < 287.6$ days

11. The rise and fall of the tides at a beach is recorded at regular intervals. Their period is almost 24 hours, but not exactly. The depth of a tidal marsh was measured over 3-hour time interval and the data is shown below.

Hours (since midnight)	0	3	6	9	12	15	18	21	24
Depth (ft)	5.5	8.0	10.5	11.7	10.8	8.4	5.8	4.3	4.9

Find a sinusoidal model for this data using your calculator. Place it in $y = a\sin(bx+c)+d$ form. Round all coefficients to the nearest *thousandth* (3 decimal places).

$y = 3.694\sin(.252x - .748) + 7.981$

According to your model, what is the period of the tides in hours? Recall that $b \cdot P = 2\pi$.

$Per = \frac{2\pi}{b} = 24.975$ hrs